Meaning shifts in distributional models: do we speak the way the world is?

Elizaveta Kuzmenko

October 2019

When we talk about objects that exist in the real world, how differently are they represented in our speech compared to their real properties as we perceive them? Which entities have similar representations in our speech but are completely different when it comes to life? In this project we want to investigate how much the world depicted in the distributional space differs from the real world. It was previously shown that the representations for colors of some entities in our speech are modified according to Gricean maxims (Rawee, 2018). Now we want to investigate these differences not only for the color subspace but on a large scale. In addition, current research in the field suggests that one semantic space can be mapped to another semantic space (Vecchi et al., 2011; Herbelot and Vecchi, 2015). However, it may be the case that two spaces differ too much, so a high quality mapping may not exist. Thus, in our research we also want to investigate how well the linguistic distributional space is mapped to the world space.

In order to perform such a comparison, we build a model representing the real world from the Visual Genome dataset (Krishna et al., 2017), a large database of images marked with objects, attributes, and relations. The language use space is built from the English Wikipedia corpus, and the British National Corpus (BNC). Exploration methods include nearest neighbor comparison and mapping frequency lists between spaces. The comparison of spaces shows that there are significant differences in concept representations between spaces, resulting in semantic shifts for the majority concepts in the language.

References

Herbelot, A. and E. M. Vecchi (2015). Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 22–32.

Krishna, R., Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision 123*(1), 32–73.

Rawee, J. (2018). The color subspace in distributional semantics: Between utterance conservation and world transformation. *Master's thesis, University of Trento*.

Vecchi, E. M., M. Baroni, and R. Zamparelli (2011). (linear) maps of the impossible: capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pp. 1–9. Association for Computational Linguistics.